Regular Article

International Journal of Toxicology
2018, Vol. 37(2) 125-143
© The Author(s) 2018
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1091581818754330
journals.sagepub.com/home/ijt
SAGE

Role of Risk of Bias in Systematic Review for Chemical Risk Assessment: A Case Study in Understanding the Relationship Between Congenital Heart Defects and Exposures to Trichloroethylene

Daniele Wikoff[1], Jon D. Urban[2], Seneca Harvey[3], and Laurie C. Haws[2]

Abstract

The National Academy of Science has recommended that a risk of bias (RoB; credibility of the link between exposure and outcome) assessment be conducted on studies that are used as primary data sources for hazard identification and dose–response assessment. Few applications of such have been conducted. Using trichloroethylene and congenital heart defects (CHDs) as a case study, we explore the role of RoB in chemical risk assessment using the National Toxicology Program's Office of Health Assessment and Translation RoB tool. Selected questions were tailored to evaluation of CHD and then applied to 12 experimental animal studies and 9 epidemiological studies. Results demonstrated that the inconsistent findings of a single animal study were likely explained by the limitations in study design assessed via RoB (eg, lack of concurrent controls, unvalidated method for assessing outcome, unreliable statistical methods, etc). Such limitations considered in the context of the body of evidence render the study not sufficiently reliable for the development of toxicity reference values. The case study highlights the utility of RoB as part of a robust risk assessment process and specifically demonstrates the role RoB can play in objectively selecting candidate data sets to develop toxicity values.

Keywords

risk of bias, systematic review, internal validity, data quality, trichloroethylene

Introduction

In recent years, there has been significant interest in integrating systematic review (SR) into toxicology and risk assessment, as doing so will aid in modernization of evidence-based decision-making.[1-5] In their recent reviews of the United States Environmental Protection Agency's (USEPA's) Integrated Risk Information System (IRIS), the National Academy of Science (NAS) recommended using SR as a means to substantially strengthen the IRIS process.[5,6] Further, the NAS[5] specifically addressed the importance of assessing the risk of bias (RoB), stating that "an ROB assessment should be conducted on studies that are used by USEPA as primary data sources for the hazard identification and dose–response assessment." That is, RoB should be evaluated for all studies used to draw conclusions regarding a potential hazard, as well as all studies used to develop toxicity values such as an oral reference dose (RfD) or reference concentration (RfC).

Numerous other investigators have identified the evaluation of "RoB" as a critical element of SR.[1,5,7,8] Assessment of the RoB involves critically appraising studies using a formal process that assesses specific aspects of quality associated with study design.[8] This process provides a measure of whether the design and conduct of a study compromised the credibility of the link between exposure and outcome.[2,7] More specifically, RoB relates to the internal validity of a study—that is, evaluation of the potential for a systematic error (ie, deviation from true effect)—that can impact the direction and magnitude of the results.[5] Assessment of RoB in SR has long been applied in the fields of medicine and other scientific disciplines; as such, many tools and frameworks exist for evaluation of RoB in clinical medicine.[7]

However, owing to both the recent application of SR in the field of toxicology[9] and the high level of heterogeneity of toxicological data sets (ie, evidence from observational human

[1] ToxStrategies, Asheville, NC, USA
[2] ToxStrategies, Austin, TX, USA
[3] ToxStrategies, Katy, TX, USA

Corresponding Author:
Daniele Wikoff, ToxStrategies, Inc, 825C Merrimon Avenue #103, Asheville, NC 28804, USA.
Email: dwikoff@toxstrategies.com

studies, experimental animal studies, and in vitro studies) relative to clinical medicine (ie, evidence primarily from controlled human trials), only 2 tools exist for the evaluation of RoB in toxicological data sets. One of the tools, which is the most relevant for chemical risk assessment, was developed by the National Toxicology Program's (NTPs) Office of Health Assessment and Translation (OHAT) and represents an RoB rating tool for both human and animal studies.[10,11] The OHAT RoB rating tool was developed for use as part of their handbook for conducting SRs. The other RoB tool that includes evaluation of animal data was developed by the SYstematic Review Centre for Laboratory animal Experimentation,[12] developed in the context of preclinical research. Both tools are based on well-established RoB guidelines developed for clinical medicine and use criteria similar to those applied to human randomized control trials, as experimental animal studies are similar in their ability to control for exposure and dose, as well as to measure outcomes. The use of the OHAT tool, which includes both human and animal studies, allows for comparison of RoB across a body of evidence, thus facilitating comparisons of data from respective evidence streams (ie, human, animal).[2] It has been recognized, however, that application of RoB tools to toxicological data sets and generation of empirical data will likely result in refinement of RoB tools and approaches as applied to toxicological data sets.

Although the conduct of RoB is clearly established as an integral component of an SR, the actual utilization of an RoB assessment in an SR supporting chemical risk assessment is less well established. Available guidance describes how to use RoB in assessing the quality in a body of evidence, but this is generally limited to evaluation of potential hazard.[1,13,14] However, it is reasonable to carry forward the concepts of study quality when selecting candidate studies (and thus carrying out the recommendations from the NAS described above). No applications of utilizing the RoB assessment to inform selection of candidate studies for development of toxicological values (such as an RfD or RfC) are available. Given the NAS recommendation to do so, and the anticipated future use of RoB in chemical risk assessment, practical applications are needed to begin establishing best practices. The need for such is highlighted by anticipated future efforts such as the USEPA's recently released *Procedures for Prioritization of Chemicals for Risk Evaluation Under the Toxic Substances Control Act, Final Rule.*[15] In the Agency's guidance document (a document designed to assist in the development of risk evaluations submitted to the USEPA under the Toxic Substances Control Act), it is recommended that a data quality system be utilized, but no additional guidance or definitions are provided.

The evidence base for trichloroethylene (TCE) provides an opportunity to explore the impact of assessing RoB in risk assessment and specifically impact on characterizing hazard and developing toxicity reference values. Although there are a number of issues that have been raised related to the evidence base related to the potential for development of congenital heart defects (CHDs) following in utero exposures to TCE,[16-19] the most notable issue concerns the selection of 1 study in particular (ie, Johnson et al[20]) for hazard characterization and development of noncancer toxicity values. This study is one of the co-candidate studies supporting the current USEPA RfD and RfC values.[21] A number of investigators have identified specific shortcomings of the Johnson et al's[20] study including issues with study design, conduct, and reporting.[16-19,22-25] Additionally, the findings reported by Johnson et al[20] are inconsistent with others in the evidence base.[17-19,22-24] However, to date, this evidence base has not been subject to a formal assessment of the RoB, nor has there been a formal assessment and integration of data quality as it pertains to developing conclusions.

Given (1) the NAS recommendations that an RoB assessment be conducted on studies used as primary data sources for the hazard identification and dose–response assessment, (2) the need for case studies and empirical evidence in testing RoB schemes for toxicological data sets, and (3) the suitability of the TCE evidence base as a case study, the objective of this current evaluation was to evaluate the RoB, as well as other data quality elements, in the evidence base considered by Makris et al[25] and to integrate such into the development of conclusions. The process implemented in this assessment followed that developed by NTP OHAT. This case study provides a demonstration as to how study quality (as evaluated by internal validity [RoB]) and external validity can be integrated into the risk assessment process, supporting both hazard characterization and the selection of candidate studies in the development of toxicity reference values.

Materials and Methods

Selection of a Case Study and Development of Evidence Base

The evidence base established by Makris et al[25] provides a readily available data set upon which to evaluate the role of RoB, as well as other elements of data quality, in chemical risk assessment. To ensure that all currently available literature was included in this RoB assessment, the evidence base developed by Makris et al[25] was combined with findings of an updated literature search (January 1, 2015, to August 15, 2017; see Supplemental Materials). The syntax was developed by an informational specialist, who also executed the PubMed and Embase searches and subsequent screening. The search strategy also involved hand searching of key primary studies as well as reviews (eg, Bukowski[26]). Additionally, while not an SR, in order to evaluate the RoB, a population, exposure, comparator, outcome (PECO) statement is required as the RoB criteria and rating instructions must be tailored to specific research questions. For the purposes of this RoB assessment, the following PECO was developed:

In humans and experimental animals, is in utero exposure to TCE associated with CHDs?. The population was defined as human and experimental animals. The exposure in question was specific to TCE, the comparator being the absence of TCE exposure (eg, control). The outcome was defined as CHDs, including defects of the valves (mitral, tricuspid, pulmonary, and aortic), arteries (aorta and pulmonary, including the transposition of major arteries), chambers (atria and ventricular), and septa (atrial, ventricular, and atrioventricular).

Critical Appraisal via RoB (Internal Validity)

A research team was assembled with expertise and experience consistent with standards for conducting RoB evaluations. Data extraction and RoB assessments were performed by 2 reviewers; conflicts were resolved by a third. Risk of bias was evaluated using the OHAT RoB tool.[11] Further, RoB was evaluated on the outcome level (vs study level) per OHAT guidance. The OHAT RoB tool is comprised of 11 questions (also known as domains) that are designed to account for different type of bias within a study that, collectively, allow reviewers to consider "the extent to which results of included studies should be relied on."[1] Each question is assigned a rating based on the following: "++" definitely low RoB (dark green shading), "+" probably low RoB (light green shading), "−" probably high RoB or not reported (light red shading), or "−−" definitely high RoB (dark red shading). The lower the RoB, the higher the methodological quality of a study/outcome.

Per guidance in using the OHAT RoB tool, it is noted that the core question of each SR is unique and therefore necessitates that investigators tailor the questions to the specific research hypothesis for a given review.[1,11] Following this guidance, 4 of the RoB questions (questions 1, 5, 8, and 9) for the experimental animal studies were evaluated by component (referred to as subdomains). That is, as written in the tool, a single question covered multiple elements of internal validity. Recognizing that part of the current objective was to evaluate RoB schemes for toxicological data sets and that some of the studies in the TCE evidence base were associated with study design limitations, it was important to be able to assess these elements separately, as well as overall. The OHAT questions differentiated by subdomain were questions 1, 5, 8, and 9 (dose randomization, identical experimental, confidence in exposure, and confidence in outcome assessment, respectively). Questions 7 and 10 were not divided into subdomains. Thus, RoB questions were evaluated as follows (see Supplemental Materials for further descriptions and rating categorizations):

Question 1a—Adequate randomization of animals to control or exposure/dose groups?

Question 1b—Were all study groups (control and exposed) investigated concurrently?

Question 5a—Was the same vehicle used for all study groups (control and exposed)?

Question 5b—Were non-treatment-related experimental conditions the same for all study groups (control and exposed)?

Question 7—Were outcome data complete without attrition or exclusion from analysis?

Question 8a—Is there confidence in test article purity?

Question 8b—Is there confidence in test agent solution concentration and stability?

Question 8c—Is there confidence that all study groups were administered doses or experienced exposures in a consistent manner?

Question 9a—Is there confidence in the outcome assessment method?

Question 9b—Is there confidence that the outcome assessors were adequately blinded to the animal/tissue study group identity?

Question 10—Were all measured outcomes reported?

Question 11—Were appropriate statistical units evaluated and reported?

In addition to customization of the criteria, OHAT also recommends that rating instructions be tailored to the specific research question. Although largely similar to that provided by OHAT, rating descriptions were refined for human and experimental animal studies, a summary of refinements are described here and details provided in the Supplemental Materials. With respect to outcome characterization for experimental animal studies, the methodology for dissection and evaluation of CHDs (question 9a) was rated for bias based on validation and reliability. Given the minute size of the fetal heart in rodents and other small animal species, and the sensitivity of this organ tissue, CHDs have been commonly identified by using 1 of 2 common and acceptable fetal dissection techniques (reviewed in Tyl and Marr[27]): the fresh in situ microdissection technique[28,29] and the fixation, serial sectioning technique.[30] Organisation for Economic Co-operation and Development (OECD) guidelines for developmental toxicity studies approve of either technique, and so both were associated with a "low" RoB for the current evaluation. There are advantages and disadvantages specific to the conduct and outcome of each method, and there is overlap in the sensitivity of each to identify certain CHDs.[31] The distinction between "definitely low" and "probably low" RoB was made based on the available evidence that indicated the "Staples technique" is overall more sensitive to the identification of malformations of the heart and major blood vessels.[27,32] Other techniques were rated based on similarity to these methods and demonstrated validation in the literature.

The 11th question, described by OHAT as "other bias," allows for additional questions for other potential threats to internal validity (eg, statistical methods) that can be added and applied as appropriate. For the experimental animal studies, the "other bias" was included, defined as, "were appropriate statistical units evaluated and reported?" For the human studies, no major modifications or subdomains were implemented. Consistent with experimental animal studies, the "other bias" question was used to account for the conduct and reporting of statistical analyses. The rating definitions were largely predicated on the appropriate use of statistical units and the handling of control groups. Because fetuses exposed in utero are wholly dependent upon the mother, and it is only the mothers who are independently sorted into study dose groups, it is a tenet of developmental toxicology that the litter—not the fetus—is the appropriate unit for statistical analysis.[27,33,34] As such, studies that reported statistical results on a per-litter basis were defined as "low" RoB for statistical analysis. Studies in which the statistical unit was not evident or was based on the fetus were defined as "high" RoB studies for this question. Further, analyses that used a single concurrent control were also considered to have lower RoB than studies that relied on pooled controls; reporting from original study reports was relied upon in assignment of rankings.

When evaluating the epidemiological literature for evidence of associations between a particular exposure birth defects, it is important to control for a number of confounding factors.[35-38] Herein, confounders considered to be important when rating epidemiological studies included maternal cigarette smoking, alcohol use, advanced maternal age, diabetes, hypertension, poor nutrition (eg, folic acid deficiency), exposure to infectious agents, and use of certain medications.[37,39-42] Particular emphasis was placed on maternal smoking, alcohol use, and hypertension, as these are factors that alone have been associated with birth defects, including CHDs.[43-47] In order to achieve a low RoB rating, epidemiology studies had to account for maternal smoking and alcohol use during pregnancy (probably low) in addition to other variables (definitely low).

Following appraisal of internal validity via RoB, studies were assigned to tiers as a means of characterizing the overall RoB for each outcome/study, thus allowing for comparison between studies and across evidence streams. Per OHAT guidance[1], a 3-tier approach was implemented, where tier 1 studies represent those studies that generally have a "low" RoB (higher level of confidence) and tier 3 studies generally have a "high" RoB (lower level of confidence). Tier 2 studies are those that met neither of the criteria for first or third tiers. Similar to that described by the OHAT guidance, the tiering approach implemented here placed emphasis on key questions. Due to the nature of experimental versus observational study types, the key questions identified for animal versus human studies differed. For the experimental animal studies, questions 5b (same nontreatment environmental conditions across groups) and 9a (method used to identify CHD) were identified as key. For the human studies, the questions identified by OHAT (4, 8, and 9) were used as key RoB domains. Tiers were defined as follows:

- Tier 1: A study must be rated as "definitely low" or "probably low" RoB for key elements and have most other applicable items answered "definitely low" or "probably low" RoB.
- Tier 2: A study that neither meets the criteria of tier 1 or tier 3.
- Tier 3: A study must be rated as "definitely high" or "probably high" RoB for key elements and have most other applicable items answered "definitely high" or "probably high" RoB.

Data Integration and Overall Evaluation of Confidence in the Body of Evidence

Data were synthesized and integrated by study type (eg, case–control/cross-sectional, and oral/inhalation), evidence stream, and overall. Confidence (also referred to as the quality of evidence) was determined per OHAT. In brief, in accordance with this guidance, an initial confidence rating is assigned based on 4 study design elements (controlled exposure, exposure prior to outcome, individual outcome data, and comparison group used). The initial confidence can then be increased based on large magnitude of effect, evidence of a dose–response, residual confounding, and consistency of results across studies. Confidence can be decreased by inconsistent results among studies, indirectness (external validity or generalizability, evaluated both on an individual study basis as well as on body of evidence basis), and imprecision. Publication bias and residual confounding were not evaluated here. Final confidence ratings were assigned by stream and overall. It should be noted, however, that the confidence ratings in the OHAT guidance reflect confidence that study findings accurately reflect the true association between exposure to a substance and effect. Thus, the framework—by default—is designed to describe confidence in observation of an effect (the alternative hypothesis) versus the lack of an effect (the null hypothesis); as such, additional narrative is required to describe confidence when data support the null hypothesis.

Evaluation of the Role and Impact of RoB on Developing Conclusions

Continuing with the OHAT process,[1,2] the confidence ratings for the body of evidence (which included consideration of RoB) were translated into evidence of health effects (step 6 in the OHAT process) and then conclusions developed based on the integration of evidence (step 7 in the OHAT process). To evaluate the potential impact of RoB, the key elements of data evaluation, including the process to do so, were considered in the context of the risk assessment process, specifically the conclusions regarding hazard and the data quality assessment relative to selection of candidate data sets, thus addressing the NAS recommendations regarding RoB assessment for studies used in dose–response assessment.

Results

Evidence Base for TCE and CHD

The literature search yielded 35 unique references published since 2015. None of the references examined the potential association of in utero exposure to TCE and development of CHDs in fetuses or neonates. Three additional epidemiological studies—Tola et al,[48] Brender et al,[49] and Gilboa et al[50]—were identified via hand searching of USEPA,[51,52] Makris et al,[25] and Bukowski.[26]

Of the 11 experimental animal studies identified, 2 reported multiple experiments (ie, evaluation of CHD in 2 different animal species).[53,54] Here, these were treated as separate studies. In addition, there were 2 publications from the same laboratory that reported on the same animal experiment conducted over a 6-year period,[20,55] as well as related correspondence and errata from the authors.[56-58] Because, collectively, these publications report on a single data set, this was treated as a single experimental animal study here and only the more recent paper[20] was included in the current RoB analysis. Similarly, for the epidemiological literature, 2 publications reported on the same investigation,[59,60] so they were evaluated as a single study. Lagakos et al[61] and Massachusetts Department of Public Health[62] also reported on the same investigation, with the latter report (published by a state government agency) presenting an updated and upgraded (cross-sectional vs cohort study) analysis of the earlier study. However, only a summary of the updated/upgraded analysis was readily available; because details were not available in such, only the earlier publication (which contained details of methods and findings) was included here.

Overall, the evidence base for TCE-CHD contained 12 experimental animal studies (Cosby and Dukelow,[63] Fisher et al,[64] Johnson et al,[20] Narotsky et al,[65] Narotsky and Kavlock,[66] Carney et al,[67] Dorfmuller et al,[68] Healy et al,[69] and 2 studies each in Hardin et al[54] and Schwetz et al[53]) and 9 epidemiology studies (Tola et al,[48] Brender et al,[49] Gilboa et al,[50] Yauck et al,[70] Bove et al[60]/Bove,[59] Forand et al,[71] Goldberg et al,[72] Ruckart et al,[73] and Lagakos et al[61]). Here, the term study refers to a unique experiment or evaluation rather than to a publication as a whole, though the author/year of a publication is used (along with a description where needed) to identify a study.

Synthesis and RoB Evaluation of Experimental Animal Studies

The TCE-CHD animal evidence base was comprised of rat (9), mouse (2), and rabbit (1) studies; these were divided into 2 groups based on route of maternal exposure (oral or inhalation; Table 1). Across the 7 inhalation studies, daily exposures to TCE ranged from 50 to 1,800 parts per million, with the exposures varying between 4 and 7 h/d over a 10- to 22-day period during gestation. With the exception of the Healy et al's[69] study (exposures in rats on gestation days 8-21), all other inhalation studies involved exposures during the critical window for fetal cardiac development (ie, gestation days 7-15, 8-13, and 8-16 for rats, mice, and rabbits, respectively).[74] No CHDs were reported in any of the TCE exposure groups in the inhalation studies, the relevant route of exposure for development of inhalation toxicity values (eg, RfC). The RoB across these studies was low to moderate; 4 studies were classified as tier 1 studies, the remaining 3 as tier 2 (Figure 1). The outcome assessment method (question 9a) is an important element of the RoB evaluation for developmental toxicity studies, given the small size and delicate nature of the fetal heart. The outcome assessments used as part of the study design for the inhalation experiments reflect common guideline methods (Staples[28] method and the close variant published by Stuckhardt and Poppe[29]; the Wilson[30] method) long recognized as appropriate for evaluating teratogenic effects in the fetuses of species used in these studies (ie, rat, mouse, rabbit), and thus, studies that used these methods were rated as "definitely" or "probably" low RoB, respectively, for question 9a. The exception was Healy et al's[69] inhalation study, which provided insufficient information on the outcome assessment methodology.

The other 5 studies involved oral exposures of pregnant mice or rats to TCE via gavage or drinking water during gestation. With the exception of the Cosby and Dukelow's[63] study (variable 5-day exposures occurring at early and mid-gestation), the windows of exposure for the oral studies ranged from 10 to 22 days and included the critical period of development for the fetal heart in rats (gestation days 7-15) and mice (gestation days 8-13).[74] Of the oral studies, only one[20] reported a statistically significant increase in CHDs in rats exposed to TCE throughout pregnancy (Table 1). Only 2 of these 5 oral studies utilized an outcome assessment recognized as a guideline method[65,66] and therefore rated a low RoB for question 9a. The remaining oral studies either provided insufficient information on the outcome methodology[63] or used a fetal heart dissection and assessment technique[20,64] that has not been validated in the scientific literature. None of the oral experimental animal studies were rated as a tier 1 study for RoB: 4 of the 5 were rated as tier 2 studies, while Johnson et al's[20] study was the only experimental animal study in the TCE-CHD evidence base to be rated as a tier 3 study (Figure 1). The Johnson et al's[20] study also had the highest RoB related to exposure characterization (question 8a-c) due to lack of information on TCE purity, failure to analytically confirm TCE concentration in daily drinking water, and exposure in a group housing setting (3 animals per cage vs individual exposures). In addition, there were a few experimental studies that had high RoB for statistical analysis (question 11) due to limitations on statistical reporting (Cosby and Dukelow,[63] Narotsky and Kavlock,[66] and Healy et al[69]) or pooling of nonconcurrent control groups (Johnson et al[20]).

Across the experimental animal evidence base, most studies had low RoB ratings for selection bias (questions 1a and b) and performance bias (ie, questions 5a and b and 7). The exception was the study by Johnson et al[20] (the only study across the evidence base to report effects), which rated high RoB for most of these subdomains. Many studies rated probably/definitely high RoB for study group concealment and blinding criteria (questions 2, 6, and 9b), as information on these elements were not reported.

Synthesis and RoB Evaluation of Epidemiological Studies

The 9 observational human studies evaluating TCE-CHDs were separated into 2 broad groups based on their level of directness (ie, external validity): (1) those that directly evaluated and reported findings specific to TCE and CHD (ie, design and report of study was "fit for purpose")[48-50,60,70,71] and (2) studies that did not evaluate or report TCE-specific exposures

Table 1. Summary of Evidence Synthesis and Confidence in the Experimental Animal Studies.

Study	Study type	Finding[a]	Included in evaluation by Makris et al (2016)?[b]	Initial confidence rating[c]	Risk of bias tier	Risk of bias	Unexplained inconsistency	Indirectness	Imprecision	Magnitude	Dose–response	Consistency across study types	Final confidence rating
Oral studies													
Cosby and Dukelow[63]	Mouse; oral gavage administered during gestation (GD 1-5, 6-10, and 11-15)	Negative: no CHDs reported in animals versus controls	Yes	High	2	–/↓ Serious (all information is from tiers 2 and 3 studies)	–Single inconsistency (1 of 5 oral study) likely explained by study design limitations. Inconsistent study could not be validated	–Not serious; direct evaluation of TCE-CHD	–/↓The single study to report effect did not report any measure of variability (SD or SE) on CHD findings	–No effects observed in 4 of 5 oral studies; single study reporting effect had low magnitude	–No TCE-CHD effects observed in 4 of 5 oral studies; single study reporting effect demonstrated poor dose–response	↑No TCE-CHD effects observed in 4 of 5 oral studies. Increases confidence in negative findings	High (+++) confidence in the animal database demonstrating null hypothesis
Fisher et al[64]	Rat; oral gavage administered during gestation (GD 6-15)	Negative: no statistically significant increase in % fetus or litter with CHDs versus controls. Positive control resulted in statistically significant increase in CHDs versus controls, validating model. Study designed to verify Dawson/Johnson positive results	Yes		2								
Johnson et al[20,d]	Rat; gestational exposure via drinking water (GD 1-22)	Positive: statistically significant increase in % fetuses with CHDs at 2 of 3 highest doses; reported a dose–response relationship	Yes		3								
Narotsky and Kavlock[66]	Rat; oral gavage administered during gestation (GD 6-19)	Negative: no CHDs reported in animals versus controls	Yes		2								
Narotsky et al[65]	Rat; oral gavage administered during gestation (GD 6-15)	Negative: no CHDs reported in animals versus controls	Yes		2								

(continued)

Table 1. (continued)

Study	Study type	Finding[a]	Included in evaluation by Makris et al (2016)?[b]	Initial confidence rating[c]	Risk of bias tier	Risk of bias	Unexplained inconsistency	Indirectness	Imprecision	Magnitude	Dose–response	Consistency across study types	Final confidence rating
Inhalation studies													
Carney et al[67]	Rat; inhalation exposure (GD 6-20, 6 h/d)	Negative: no CHDs reported in animals versus controls	Yes		1	–Not likely (Most information is from tier 1 studies)	–No inconsistency between inhalation studies to explain	–Not serious; direct evaluation of TCE-CHD	–No CHDs reported in any of the 7 inhalation studies	–No effects observed in 7 of 7 inhalation studies	–No TCE-CHD response reported in 7 of 7 inhalation studies	↑No TCE-CHD response reported in 7 of 7 inhalation studies. Increases confidence in negative findings	High (+++) confidence in the animal database demonstrating null hypothesis
Dorfmueller et al[68]	Rat; inhalation exposure (GD 1-20, 6 h/d)	Negative: no CHDs reported in animals versus controls	Yes		1								
Hardin et al[54,e]—(1) rat experiment	Rat; inhalation exposure (GD 0-18, 7 h/d)	Negative: no CHDs reported in animals versus controls	Yes		2								
Hardin et al[54,e]—(2) rabbit experiment	Rabbit; inhalation exposure (GD 0-21, 7 h/d)	Negative: no CHDs reported in animals versus controls	Yes		2								
Healy et al[69]	Rat; inhalation exposure (GD 8-21, 4 h/d)	Negative: no CHDs reported in animals versus controls	Yes		2								
Schwetz et al[53]—(1) rat experiment	Rat; inhalation exposure (GD 6-15, 7 h/d)	Negative: no CHDs reported in animals versus controls	Yes		1								
Schwetz et al[53]—(2) mouse experiment	Mouse; inhalation exposure (GD 6-15, 7 h/d)	Negative: no CHDs reported in animals versus controls	Yes		1								

Abbreviations: CHD, congenital heart defect; OHAT, Office of Health Assessment and Translation; TCE, trichloroethylene; GD, gestation day; SE, standard error; SD, standard deviation.

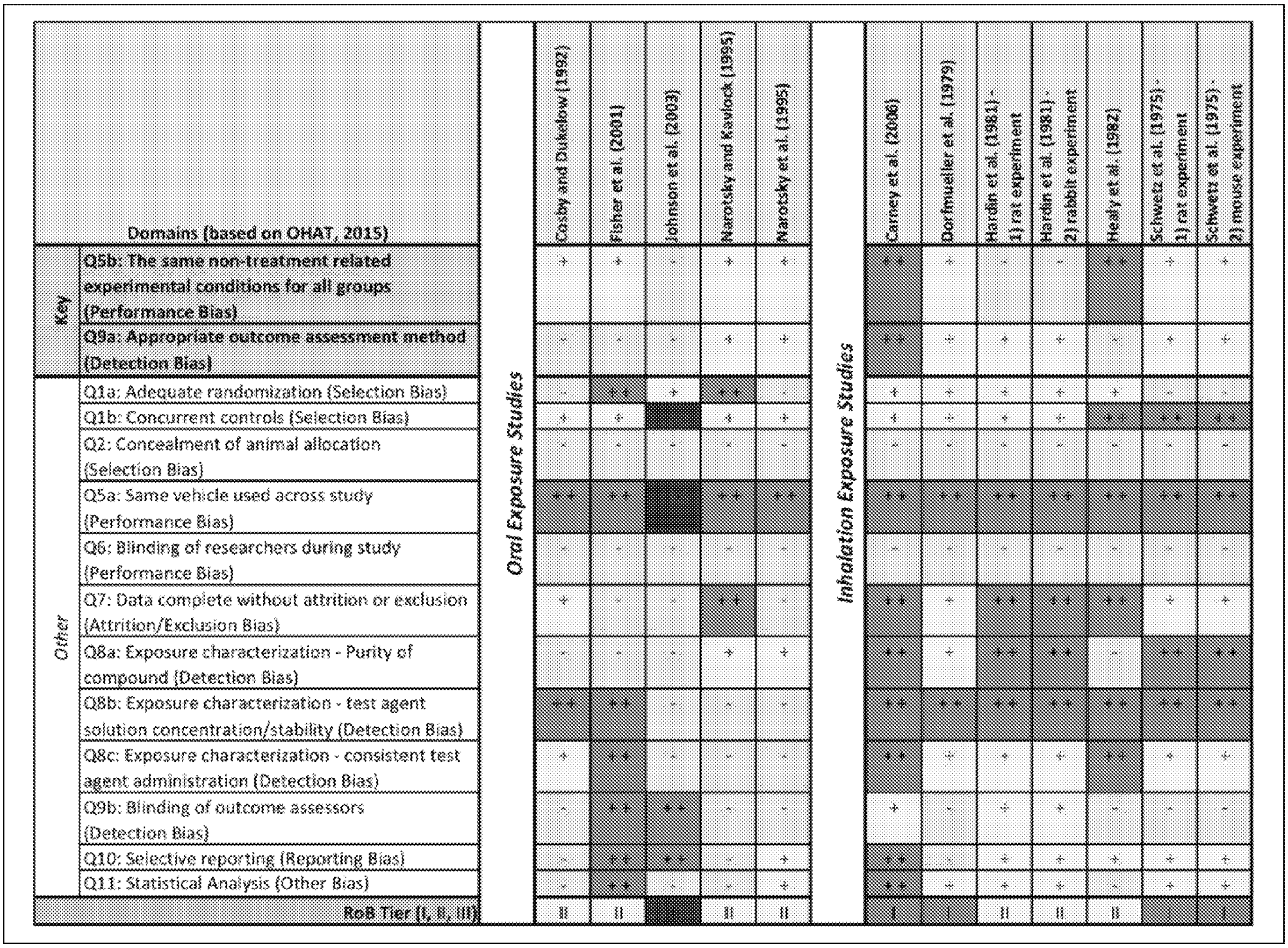
[a] Based on assessor classification; unless the study authors reported a statistically significant increase in CHDs relative to control group, finding is negative for TCE-CHD.

[b] Classification based on inclusion in evaluation of hazard (ie, inclusion of tables, etc).

[c] Based on OHAT[1]: Table 8—Study design features for initial confidence rating.

[d] As ultimately acknowledged by the original study authors, Johnson et al[20] represents multiple experiments conducted over a 6-year period.[57,58] Data presented in an earlier publication represent the results of 2 of 4 total exposure groups of this study and were republished by Johnson et al.[20] Although previous reviewers have referred to these publications as 2 distinct efforts, it is more accurate to treat these articles as a single study unit. Since the data for all 4 exposure groups were published by Johnson et al,[20] this is the citation used in this study to represent this study. However, there are certain inconsistencies between the 2 publications (vehicle used, number of controls, CHD classification) that are important to characterizing study quality, and therefore, we include selective references to the earlier publication.

[e] Hardin et al's[54] study is a general summary of a series of teratogenicity studies on several workplace chemicals performed in multiple animal species by a contract research laboratory (Litton Bionetics) on behalf of the National Institute of for Occupational Safety and Health (NIOSH). Critical study design and results information were not reported by Hardin et al[54] but were found in the publically available laboratory report furnished to NIOSH by Litton Bionetics.[92]



	Domains (based on OHAT, 2015)	Oral Exposure Studies					Inhalation Exposure Studies						
		Cosby and Dukelow (1992)	Fisher et al. (2001)	Johnson et al. (2003)	Narotsky and Kavlock (1995)	Narotsky et al. (1995)	Carney et al. (2006)	Dorfmueller et al. (1979)	Hardin et al. (1981) - 1) rat experiment	Hardin et al. (1981) - 2) rabbit experiment	Healy et al. (1982)	Schwetz et al. (1975) - 1) rat experiment	Schwetz et al. (1975) - 2) mouse experiment
Key	Q5b: The same non-treatment related experimental conditions for all groups (Performance Bias)	+	+	-	+	+	++	+	-	-	++	+	+
	Q9a: Appropriate outcome assessment method (Detection Bias)	-	-	-	+	+	++	+	+	+	-	+	+
Other	Q1a: Adequate randomization (Selection Bias)	-	++	+	++	-	+	+	+	+	+	-	-
	Q1b: Concurrent controls (Selection Bias)	+	+		+	+	+	+	+	+	++	++	++
	Q2: Concealment of animal allocation (Selection Bias)	-	-	-	-	-	-	-	-	-	-	-	-
	Q5a: Same vehicle used across study (Performance Bias)	++	++		++	++	++	++	++	++	++	++	++
	Q6: Blinding of researchers during study (Performance Bias)	-	-	-	-	-	-	-	-	-	-	-	-
	Q7: Data complete without attrition or exclusion (Attrition/Exclusion Bias)	+	-	-	++	-	++	+	++	++	++	+	+
	Q8a: Exposure characterization - Purity of compound (Detection Bias)	-	-	-	+	+	++	+	++	++	-	++	++
	Q8b: Exposure characterization - test agent solution concentration/stability (Detection Bias)	++	++	-	-	-	++	++	++	++	++	++	++
	Q8c: Exposure characterization - consistent test agent administration (Detection Bias)	+	++	-	-	-	++	+	+	+	++	+	+
	Q9b: Blinding of outcome assessors (Detection Bias)	-	++	++	-	-	+	-	+	+	-	-	-
	Q10: Selective reporting (Reporting Bias)	-	++	++	-	+	++	-	+	+	+	+	+
	Q11: Statistical Analysis (Other Bias)	-	++	-	-	+	++	+	+	+	-	+	+
	RoB Tier (I, II, III)	II	II		II	II	I	I	II	II	II	I	I

Figure 1. Risk of bias (RoB) heat map for experimental animal studies. The question-based validity was evaluated using the Office of Health Assessment and Translation (OHAT) RoB tool. Risk of bias for each question is indicated by color: "definitely low RoB" (dark green, ++), "probably low RoB" (light green, +), "probably high RoB" (light red, −), and "definitely high RoB" (dark red, −−).

or effects but were included in the evidence base by Makris et al[25] or Bukowski[26] (Table 2). These latter studies involved exposure to media that may have contained TCE or a mixture of TCE and other compounds, but authors did not attempt, or did not attribute, exposures and/or effects to TCE specifically.[61,72,73] Additionally, the information presented in the study by Goldberg et al,[72] Lagakos et al,[61] and Ruckart et al[73] showed evidence of coexposures to other chemicals (some of which, such as lead, are known to be associated with CHDs[75]). And while coexposure is evaluated in RoB, these studies were substantially different than the studies determined to be more "fit for purpose." As such, these studies were also evaluated for RoB, but as a second group, and integrated separately from the first group of studies.

The first group of studies was selected as the primary evidence base evaluating associations between TCE exposure and CHDs in humans and was comprised of 6 studies: a single cohort study (Tola et al[48]), 2 cross-sectional studies (Bove[59]/Bove et al,[60] Forand et al[71]), and 3 case–control studies (Yauck et al,[70] Gilboa et al,[50] and Brender et al,[49]). The findings from these are mixed; several of the studies report a lack of association, whereas others report weak findings for some types of malformations (but not others; Table 2). Interpretation of these data is difficult, given the heterogeneity of study design and conduct and seriousness of RoB (Figure 2). For example, Bove[59]/Bove et al[60] report an odds ratio (OR) of 1.24 for the association between TCE concentrations of >10 parts per billion (ppb) in residential wells and major cardiac effects. Interpretation is severely limited by (1) no confidence interval (CI) derived/provided by the authors, (2) lack of confidence in exposure (based on a series of assumptions relating biannual measurements of TCE in public water systems to residential status), and (3) lack of adjustment for critical confounding variables. The largest magnitude of effect was reported by Forand et al,[71] reporting an RR of 4.91 (95% CI: 1.58-15.24); however, this risk ratio estimate lacked precision, nor did it reflect an adjusted value that accounted for confounding. Additionally, this study utilized population-based exposure estimates of exposure, as opposed to exposure estimates for the individuals in the study.

Table 2. Summary of Evidence Synthesis and Confidence in Human Studies.

Study	Study type	Finding[a]	Included in evaluation by Makris et al?[b]	Initial confidence rating[c]	Risk of bias tier	Risk of bias	Unexplained inconsistency	Indirectness	Imprecision	Magnitude	Consistency across study types	Final confidence rating
Analyses involving direct assessment of TCE and CHD (ie, offered evaluation and results specific to TCE and CHD)												
Tola et al[48]	Cohort	Negative: no malformed baby was found to have been born to any mother exposed occupationally	No	Moderate	2	–/↓Serious (all information is from tier 2 studies)	–Inconsistencies assumed to be inherent to study design elements	–Not serious; direct evaluation of TCE and CHD	–When provided, confidence interval ranges varied, some were large	–When effect observed, magnitude was not large	–Results are not consistent between study types	Low (++) to very low (+) confidence in the human database demonstrating either null or alternative hypothesis
Gilboa et al[50]	Case–control	Negative: no significant increase in CHDs in mothers occupationally exposed to TCE between cases and controls ($P = 0.67$).	No	Low to moderate	2							
Yauck et al[70]	Case–control	Negative: no significant increase in CHDs between cases and controls when only distance from TCE emitters is evaluated.Positive: significant increase in CHDs between cases and controls when distance from TCE emitters is adjusted by age (adjusted OR, 3.2; 95% CI, 1.2-8.7)	Yes	Low to very low	2							
Brender et al[49]	Case–control	Negative: no increase in conotruncal heart defects (OR = 0.98; 95% CI 0.87-1.10) or obstructive heart defects (OR = 1.03; 95% CI 0.92-1.15) for cases where maternal residence was proximal to TCE source.Positive: small increase in septal heart defects (OR = 1.06; 95% CI 1.02-1.10) for cases where maternal residence was proximal to TCE air emissions (based on TRI data)	No		2							
Bove et al[59,60,d]	Cross-sectional	Weak/unclear (estimated OR with no CI): increase in major cardiac effects for TCE >10 ppb (OR = 1.24) and increase in ventral septal defects (VSDs; OR for TCE >5 ppb = 1.30), but no CI provided in either case, so significance is unclear	Yes (also combined)		2							
Forand et al[71]	Cross-sectional	Positive: increase for all CHDs (RR = 2.15; 95% CI: 1.27-3.62); major cardiac defects (n = 6) were of borderline statistical significance (RR = 2.40; 95% CI: 1.00-5.77), but significant increase in conotruncal defects (RR = 4.91; 95% CI: 1.58-15.24; authors note that although statistically significant, these RRs were	Yes		2							

(continued)

Table 2. (continued)

Study	Study type	Finding[a]	Included in evaluation by Makris et al?[b]	Initial confidence rating[c]	Risk of bias tier	Risk of bias	Unexplained inconsistency	Indirectness	Imprecision	Magnitude	Consistency across study types	Final confidence rating
		based on 3 and 4 exposed cases, respectively										
Analyses involving assessment of exposure to media with multiple contaminants including TCE (ie, no evaluation and results specific to TCE)[e]												
Goldberg et al[72]	Case–control (pseudo)	Weak/unclear (estimated OR with no CI): CHDs for exposure to water contaminated with TCE and other chemicals	Yes	Low (was not a true case–control study)	3	↓Very serious: plausible bias that seriously weakens confidence in the results (2 of 3 studies rated tier 3)	–Inconsistencies assumed to be inherent to study design elements	↓Serious: CHD not directly assessed with potential TCE exposure, but more general contaminated water source	↓Precision cannot be evaluated since no measures of variability were provided	–Two of 3 studies reported no effect; when effect observed, magnitude was not large	–Results not adequately analyzed or reported to evaluate consistency across studies	Very low (+) confidence in the human database demonstrating either null or alternative hypothesis
Ruckart et al[73]	Case–control	Negative: CHDs in TCE-contaminated area of Camp Lejeune reported only one-third expected number. Authors did not analyze CHD data further	Yes	Low to moderate	3							
Lagakos et al[61]	Cross-sectional	Negative: No results for TCE specifically; negative CHDs for exposure to well water contaminated with TCE and other chemicals	Yes	Low to very low	2							

Abbreviations: CHD, congenital heart defect; CI, confidence interval; OR, odds ratio; TCE, trichloroethylene; TRI, USEPA's Toxic Release Inventory program; RR, relative risk.

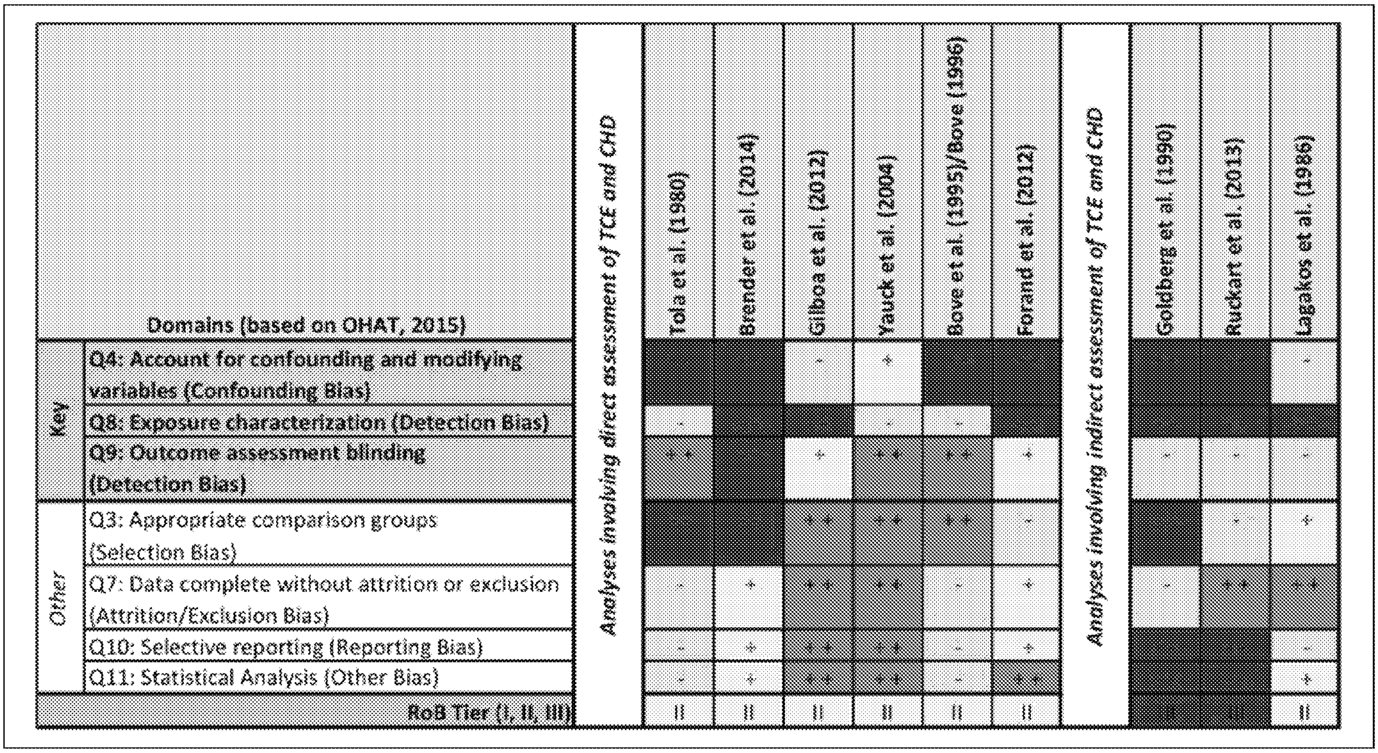
[a] Based on assessor classification; OR with CIs overlapping were not considered to be positive.

[b] Classification based on inclusion in evaluation of hazard (ie, inclusion of tables, etc); Gilboa et al[50] and Brender et al[49] are cited in the narrative but not formally included in the evaluation.

[c] Based on Office of Health Assessment and Translation[1]: Table 8—Study design features for initial confidence rating.

[d] Both studies report on the same assessment and provide similar information and thus are regarded as a single line of evidence.

[e] Studies that were included by Makris et al[25] as part of the evaluation of TCE, but authors do not provide TCE-specific analyses; other studies characterizing CHD from media containing chlorinated contaminants or other agents are available but not included in this study (ie, the scope involved TCE and CHD, based on evaluation by Makris et al[25]).



	Domains (based on OHAT, 2015)	Analyses involving direct assessment of TCE and CHD	Tola et al. (1980)	Brender et al. (2014)	Gilboa et al. (2012)	Yauck et al. (2004)	Bove et al. (1995)/Bove (1996)	Forand et al. (2012)	Analyses involving indirect assessment of TCE and CHD	Goldberg et al. (1990)	Ruckart et al. (2013)	Lagakos et al. (1986)
Key	Q4: Account for confounding and modifying variables (Confounding Bias)		■	■	-	+	■	■		■	■	-
	Q8: Exposure characterization (Detection Bias)		-	■	■	-	-	■		■	■	■
	Q9: Outcome assessment blinding (Detection Bias)		++	■	+	++	++	+		-	-	-
Other	Q3: Appropriate comparison groups (Selection Bias)		■	■	++	++	++	-		■	-	+
	Q7: Data complete without attrition or exclusion (Attrition/Exclusion Bias)		-	+	++	++	-	+		-	++	++
	Q10: Selective reporting (Reporting Bias)		-	+	++	++	-	+		■	■	-
	Q11: Statistical Analysis (Other Bias)		-	+	++	++	-	++		■	■	+
RoB Tier (I, II, III)			II	II	II	II	II	II		[illegible]	[illegible]	II

Figure 2. Risk of bias (RoB) heat map for epidemiological studies. The question-based validity was evaluated using the Office of Health Assessment and Translation (OHAT) RoB tool. Risk of bias for each question is indicated by color: "definitely low RoB" (dark green, ++), "probably low RoB" (light green, +), "probably high RoB" (light red, –), and "definitely high RoB" (dark red, ––).

The study with the lowest overall RoB, Yauck et al,[70] reported a lack of association for TCE when unadjusted for potential confounders but reported an increased OR when adjusted for certain risk factors (3.2; 95% CI: 1.2-8.7). This case–control study was the only study in the evidence base that adjusted for both maternal smoking and alcohol consumption—variables that the authors found to be significant on their own,[70] thus highlighting the critical nature of evaluating such. The study by Gilboa et al,[50] a case–control study that evaluated occupational exposures to TCE (and other solvents) in women from the National Birth Defects Prevention Study, did not find a significant increase in CHDs between cases and controls ($P = 0.67$). Notably, the study by Gilboa et al[50] was the only study in the evidence base to adjust for folic acid supplementation, although the authors did not adjust for alcohol consumption or smoking patterns. As demonstrated in Figure 2, adjustment for confounding was a significant limitation across the evidence base.

More significant than confounding, however, are the limitations in evaluation of exposure across the evidence base. None of the studies directly measured exposure in subjects; this is a critical limitation as such studies are likely to have less RoB than studies involving indirect measures. Two studies utilized proximity to a TCE source as a measure of exposure,[49,70] 2 used group-level categorical classifications based on residential location,[59,71] and 2 used occupational status, either via job exposure matrix (nonvalidated and based on self-reporting, thus introducing the potential for recall bias)[50] or via biomonitoring data (urinary trichloroacetic acid).[48] Using proximity as a surrogate for exposure, rather than using analytical data to model exposure estimates, is known to produce biased results.[76] The utilization of proximity to exposure sources greatly reduces the available information and introduces sources of bias, both mathematically and with respect to researchers' judgment. In the absence of an analysis of the various distances that comprise a study's data set, this also suggests some significant relations could only be detected using the selected bands of distance (eg, living within 1.32 miles of at least 1 site, as was categorically evaluated by Yauck et al[70]; use of a "threshold distance" (undefined) by Brender et al[49]), which casts doubt on the validity of the findings. If living near these sites were associated with higher risk, using the continuous number of sites nearby or several continuous variables documenting continuous distance to the nearest 3 sites or simply using the geographical coordinates of the households versus exposed/nonexposed categorization based on a specific distance (eg, 1.32 miles) would also eliminate some of the bias and lend credibility to the findings.

Additionally, OHAT includes verification of the compound over the course of the test period as an element in determining exposure misclassification, underscoring the importance of accounting for changes in media levels of volatile compounds during the course of the study.[11] Only 1 study in the human evidence base involved direct measurement of TCE in any form—Bove[59]/Bove et al[60] The authors of this study utilized data from biannual measurements of TCE in drinking water.

Given the volatility of TCE, there is low confidence that biannual measurements represented an accurate characterization of exposures to TCE via the public water supply.

Moreover, with the exceptions of the studies by Brender et al[49] and Bove[59]/Bove et al,[60] none of the studies adjusted risk estimates for the potential impact of coexposure to other chemicals on the TCE-CHD association data. This limitation is of particular relevance to 3 of the studies that were categorized separately due to lack of TCE-specific evaluation and reporting. The studies Lagakos et al,[61] Goldberg et al,[72] and Ruckart et al,[73] all involve exposure to media with multiple contaminants (eg, dichloroethylene, tetrachloroethylene, chloroform, lead, chromium, etc; direct evidence of such provided by study authors). Two of these studies reported a lack of CHD response in their respective study populations: Lagakos et al,[61] using a space–time distribution from wells and survey data of adverse pregnancy outcomes, and Ruckart et al,[73] in an evaluation of birth defects in babies born to women who lived on Camp Lejeune during their pregnancy. The CHD findings in the latter study are only presented as part of the methods, with the authors reporting that less than the expected number of cases of conotruncal heart defects were observed in the Camp Lejeune population, which the authors provide as justification for excluding CHDs from their agent-specific assessments. Additionally, both of these studies relied upon self-reporting of outcome (and thus the potential for recall bias exists). It should be noted that Lagakos et al[61] attempted to check the accuracy of the outcomes via medical confirmation, findings of which suggested a low rate of false positives, and that overreporting was infrequent and not more common among exposed respondents. The third study—a nontraditional case–control study published by Goldberg et al[72]—reported a relative OR that was "3 times greater" (actual OR not provided) based on comparisons of exposed and unexposed cases (a comparison associated with a high RoB). As a group, these 3 studies had a high RoB for most questions relevant to human studies, including all 3 of the key questions (ie, confounding [eg, no evaluation of confounding], exposure [eg, residence and/or estimation of the fraction of water from selected wells], and outcome evaluation [eg, self-report from telephone survey]; Figure 2).

Evidence Integration and Confidence in Body of Evidence

Tables 1 and 2 summarize the elements of evidence integration and resulting confidence in the body of evidence for TCE and CHDs as evaluated per NTP[1] for the animal and human evidence streams, respectively. The experimental animal studies had an overall lower RoB (mostly tier 1/2 and a single tier 3) than the human data (mostly tier 2 of 3 studies). For the experimental animal data, both oral and inhalation studies were assigned initial confidence ratings of "high," per NTP.[1] Findings of the inhalation studies were consistent (all 7 studies resulted in the same result, lack of effects). Collectively, these inhalation studies were considered "not likely" to have significant RoB, a low level of indirectness (ie, high-level confidence in the external validity or generalizability of these data), and no unexplained inconsistencies. And thus, the final level confidence in the studies was very high, that is, there is a very high level of confidence in the evidence base supporting a lack of association between inhalation of TCE and CHDs in experimental animal studies.

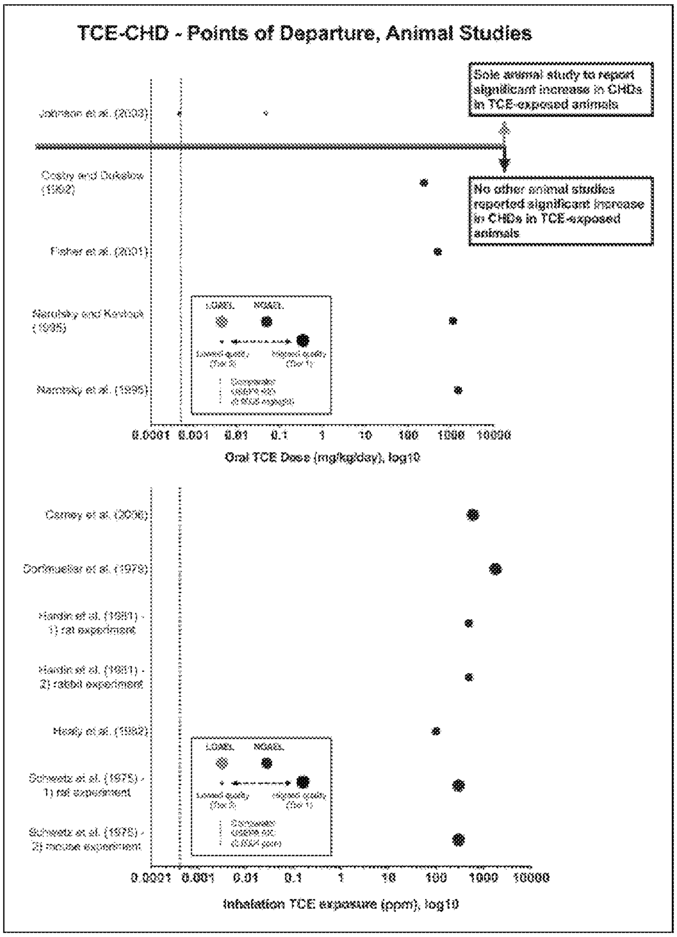
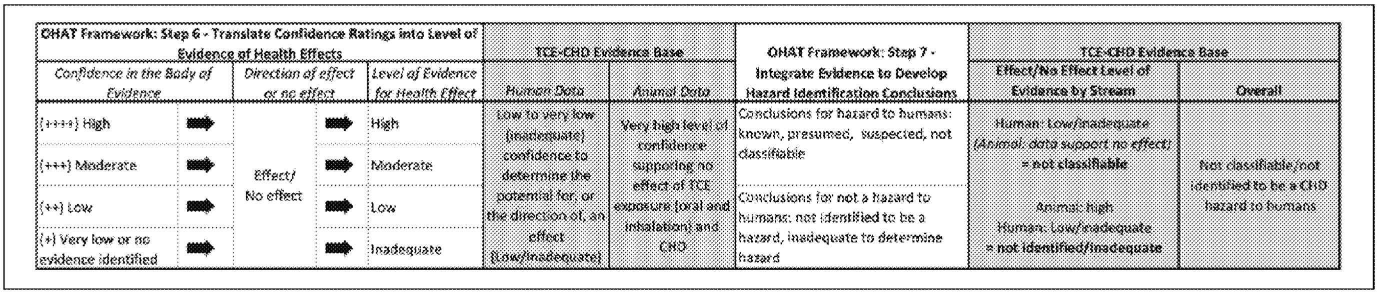


Figure 3. Summary diagram of exposure–response data for trichloroethylene (TCE) exposure via oral (A) or via an inhalation route (B) and congenital heart defects in experimental animal studies. Symbols represent intake dose as reported by original study authors. The color of the symbol indicates the type of effect: no observed adverse effect level (NOAEL; blue symbols) or the lowest observed adverse effect level (LOAEL; orange symbols). The size of the symbol indicates the overall risk of bias (ie, larger symbols indicate a lower risk of bias—or higher methodological quality, and vice versa). The dashed vertical line marks current United States Environmental Protection Agency (USEPA) reference concentration (RfC, A) and RfC (B).

A similar final level of confidence was determined for the experimental animal studies involving oral exposure. Only 1 of the 5 oral studies reported CHDs following in utero exposure to TCE (Figure 3). This finding, which is inconsistent with all other oral studies, is explained by high risk of performance, detection, selection, and other (statistical) bias, specifically the lack of concurrent controls, lack of consistent vehicles across control and dose groups, uncertainty in exposures, use of unique and unvalidated outcome assessment method, and



<table>
<tr><th colspan="5">OHAT Framework: Step 6 - Translate Confidence Ratings into Level of Evidence of Health Effects</th><th colspan="2">TCE-CHD Evidence Base</th><th rowspan="2">OHAT Framework: Step 7 - Integrate Evidence to Develop Hazard Identification Conclusions</th><th colspan="2">TCE-CHD Evidence Base</th></tr>
<tr><th colspan="2"><i>Confidence in the Body of Evidence</i></th><th colspan="2"><i>Direction of effect or no effect</i></th><th><i>Level of Evidence for Health Effect</i></th><th><i>Human Data</i></th><th><i>Animal Data</i></th><th>Effect/No Effect Level of Evidence by Stream</th><th>Overall</th></tr>
<tr><td>(++++) High</td><td>➡</td><td rowspan="4">Effect/
No effect</td><td>➡</td><td>High</td><td rowspan="4">Low to very low (inadequate) confidence to determine the potential for, or the direction of, an effect (Low/inadequate)</td><td rowspan="4">Very high level of confidence supporing no effect of TCE exposure (oral and inhalation) and CHD</td><td rowspan="2">Conclusions for hazard to humans: known, presumed, suspected, not classifiable</td><td rowspan="4">Human: Low/inadequate
<i>(Animal: data support no effect)</i>
= not classifiable
Animal: high
Human: Low/inadequate
= not identified/inadequate</td><td rowspan="4">Not classifiable/not identified to be a CHD hazard to humans</td></tr>
<tr><td>(+++) Moderate</td><td>➡</td><td>➡</td><td>Moderate</td></tr>
<tr><td>(++) Low</td><td>➡</td><td>➡</td><td>Low</td><td rowspan="2">Conclusions for not a hazard to humans: not identified to be a hazard, inadequate to determine hazard</td></tr>
<tr><td>(+) Very low or no evidence identified</td><td>➡</td><td>➡</td><td>Inadequate</td></tr>
</table>

Figure 4. Application of the Office of Health Assessment and Translation (OHAT) framework for systematic review and evidence integration for developing hazard identification conclusions (steps 6 and 7).

pooling of nonconcurrent control group data. When compared to other studies with lower RoB (ie, concurrent controls, consistent vehicle across groups, analytical certainty of exposure dose/exposure levels, common and validated outcome assessment methods, and appropriate statistical analyses) that evaluated similar and higher exposure doses and exposure paradigms, it is apparent that the Johnson et al's[20] study is not sufficiently reliable for hazard characterization or for development of noncancer toxicity values. This is further supported by the lack of ability to replicate the study's findings in a study designed specifically to do so (Fisher et al[64]; particularly notable given that the first author of the Johnson et al's[20] study was also as a member of the cardiac dissection and assessment team in the study by Fisher et al[64]).

For the human studies, initial confidence ratings based on study type ranged from moderate to very low. When the "serious" and/or "very serious" RoB was considered along with inconsistent findings, imprecision, and low magnitude of effects, there was an overall decrease in confidence. That is, there is a very low to low level of confidence in the body of evidence. That is, there are no data of sufficient quality (available data have low to very low level of confidence) to determine the direction of an effect (consistent with OHAT methodology, evidence receiving "very low" confidence ratings should not be used to develop conclusions regarding the potential for health effects; OHAT and Rooney et al[2]).

Integrated Conclusions Considering RoB

Per the OHAT framework, the RoB assessment and level of confidence ratings (steps 4 and 5 in the OHAT framework) were carried forward to the development of conclusions. This involved translating confidence ratings into levels of evidence for health effects (step 6) and classification of overall conclusions (step 7). For the human evidence base, the confidence ratings translated into a "low to inadequate" level of evidence, that is, there is a low to very low (inadequate/insufficient) confidence to determine the potential for, or the direction of, an effect of TCE exposure and CHDs. For the animal evidence base, recognizing that the single inconsistency can be explained by study design, conduct, and reporting limitations, it was determined that the final confidence rating for the oral studies was "high." That is, there was a high level of confidence supporting a lack of association between oral or inhalation exposure to TCE and CHDs in experimental animal studies. In making this determination, contextual (confirmatory) efforts related to the sensitivity of the experimental animal studies were also considered; unlike known cardioteratogens (eg, alcohol, retinoic acid), the animal and human in utero exposure studies provide no evidence of any particular CHD pattern or predominant CHD associated with TCE exposure.

The translated levels of evidence for each stream were then integrated using the matrix provided by OHAT. Per OHAT methodology, data receiving a "very low" level of confidence rating or an "inadequate" level of evidence do not move forward to the development of conclusions; in such cases, it is recommended that conclusions are based on the remaining evidence stream alone. The TCE-CHD evidence base is difficult to integrate, given the lack of confidence to determine the potential for, or direction of, an effect in the human data. Using a conservative approach, and assuming a low (vs inadequate) level of effect for the human data, combined with the high level of confidence that TCE is not associated with CHDs in animals, the overall conclusion ranges from classification of TCE as "not classifiable" to "not identified" to be a CHD hazard (Figure 4).

Impact of RoB

In the context of risk assessment, the resulting impact of the RoB assessment on TCE-CHD is the determination that CHDs are not the most suitable end point upon which to base a quantitative assessment and that the Johnson et al's[20] study is not sufficiently reliable for hazard characterization or development of noncancer toxicity values.

Discussion

The RoB assessment described here provided a systematic, transparent approach to evaluating methodological quality. Following NAS recommendations to conduct an RoB assessment on studies used as primary data sources for dose–response assessment, we have demonstrated that one of the co-candidate studies used to develop the current RfD and RfC values for TCE has the highest RoB in the evidence base. Further, this case study demonstrates that the inconsistent finding of this study (Johnson et al[20]) could be explained by bias in selection, performance, detection, exposure, and statistics (eg, lack of concurrent controls, lack of consistent vehicle between control and exposure groups, uncertain exposure levels in TCE-exposed animals, unvalidated method for assessing outcome, unreliable statistics, etc). Due to the high RoB (tier 3), inconsistent findings with all other animal studies ($n = 11$, all of which had lower RoB ratings) and inability to replicate study findings, results of this case study demonstrated that the Johnson et al's[20] study is not sufficiently reliable for hazard characterization or development of noncancer toxicity values. And thus, using the process described here regarding the role of RoB in selecting reliable candidate studies to serve as the basis of toxicity values, the literature characterizing other end points (including alternative developmental effects) could be evaluated and a more reliable and representative data set (or data sets) selected.

The RoB evaluation conducted here demonstrates the importance of evaluating and integrating RoB both in developing hazard conclusions and in candidate data selection for dose–response assessment and development of toxicity values. It also highlights the significant utility of implementing an SR process (such as that described by OHAT process) in risk assessment. Based on decades of experience from the fields of toxicological and clinical medicine, the OHAT approach provides a transparent, objective process for characterizing the validity of the evidence, rating confidence in the evidence, translating confidence in the body of evidence to level of evidence in health effects, and finally to integrating the evidence in developing hazard identification conclusions. Thus, individual study quality is inherent to the synthesis and development of conclusions. Moreover, the OHAT approach guides the user to make conclusions on reliable data, and if such are not available, to be transparent in classifications, utilizing terms such as "insufficient," "inadequate," or "not classifiable" (ie, weak or low levels of evidence between streams do not relate to a high level of evidence of effect).

The OHAT approach, however, is limited to hazard classifications. As demonstrated here, the output of an SR can readily be utilized in subsequent steps in a risk assessment. The particular utility of carrying the output forward is demonstrated via comparison of this case study with a review on a similar body of evidence that did not include an assessment of the RoB,[25] which resulted in an opposite conclusion regarding the suitability of the Johnson et al's[20] study for development of noncancer toxicity . Differences in the conclusions can be explained by elements of the RoB assessment. For example, an RoB assessment is conducted at the outcome (vs study) level. As such, the publications by Dawson et al[55] and Johnson et al[20] (and associated errata) were handled as a single experimental study in this case study, since the data set in Johnson et al[20] includes all the TCE-CHD data from the earlier paper. In contrast, Makris et al[25] treats these studies inconsistently, considering them separate and independent studies for much of their assessment (which gives the perception of a greater volume of evidence than is actually available), but as a single study for the dose–response evaluation. The question-based evaluation of RoB conducted here provided an objective rationale for assessment of internal validity—the output of which transparently provides rationale for the lack of reproducibility, low magnitude of response, and the likely reasons for the inconsistency in findings (ie, performance, detection, and selection biases). In this case study, both the findings and the study quality (as assessed by internal and external validity) of all of the evidence were integrated, whereas Makris et al[25] did not formally integrate the studies reporting a lack of TCE-CHD association in rats, mice, and rabbits.[54,63,65,66,68,69]

In making these comparisons, it is notable that evaluation and integration of RoB did not result in significantly different conclusions from Makris et al[25] regarding the human studies despite differences in overall approach. It is likely that similar conclusions were reached for the human evidence because (1) some aspects of bias were considered (though not formally evaluated) by Makris et al[25] and (2) there is overlap in the weight of the evidence approach used by Makris et al[25] and the elements that also form the basis of Grading of Recommendations Assessment, Development, and Evaluation (GRADE)and OHAT evidence integration frameworks. For example, Makris et al[25] informally considered confounding variables, approach for evaluation exposure, and classification of outcomes. The general conclusion on the lack of reliability of the available human evidence is consistent with that of prior reviews of this literature (eg, Hardin et al,[54] Watson et al,[18] and Makris et al[25]). The RoB conducted here also contributes to an additional need identified by Makris et al[25] regarding interpretation of the epidemiological database for cardiac defects associated with TCE exposures. Presently, the high level of heterogeneity in study design and the lack of information within individual studies (ie, no OR developed, no CIs reported) preclude meta-analyses.

The findings of the case study reinforce the OHAT recommendation regarding a priori project-specific customization of the RoB approach to rigorously evaluate and differentiate study quality for a given PECO. For example, here, we identified and categorized outcome assessment methods associated with the lowest RoB for cardiac heart defects in experimental animal studies. This was based on the classification of dissection methods used in OECD guidelines (or similar) as having a low RoB. Doing so allowed for further differentiation of study quality (an objective of the assessment). The majority of TCE-CHD studies used guideline-approved dissection methods. Two studies used a dissection technique that was not considered to be reliable here: Johnson et al[20] and Fisher et al,[64] the latter of which was explicitly designed to attempt to replicate the CHD findings from Johnson et al.[20] Dawson et al[55] described this alternative dissection technique and alleged that it was sensitive to the detection of particular defects (eg, adhered valve cusps) and abnormal valve dimensions (Johnson et al[77]). It should be noted that the controls in these 2 studies also had considerably higher background levels of CHDs relative to the Staples technique (Carney et al[67]). This suggests that the combination of the fixing and unique tissue cuts on such minute tissues may be introducing artifacts. As such, the dissection method used in these 2 studies (Fisher et al[64] and Johnson et al[20]) was not considered to be reliable. It is also recognized, however, that the types of CHDs reported in these studies were diverse and inconsistent among TCE treatment groups, with no evidence of a predominant defect or set of defects in any TCE exposure group in these studies.[18,19,51] A similar situation arises when evaluating the CHD data presented in the TCE metabolite studies.[78-81]

Implementation of the case study also reinforced that an RoB assessment does not eliminate subjectivity and expert judgment, though highlighting the complimentary nature of utilizing a transparent, formal system to evaluate RoB and integration of such in decision-making. For example, when evaluating the potential for bias, this current evaluation differed from Makris et al[25] as to what would constitute bias selection and performance bias, specifically with respect to what constitutes an appropriate control group. Makris et al[25] considered the pooling of 5 groups of nonconcurrent control animals that received different vehicles to be analogous to a historical control group and thus suitable for use as a control in the statistical analyses. Makris et al[25] further characterized this heterogeneous combination of data across studies as a strength. In contrast, here, these factors were viewed as shortcomings in methodological quality, relating to a high RoB in several questions. It is also notable that in recognizing some of these aspects as potential shortcomings, Makris et al[25] contacted the original study authors for clarification and cite personal communications in which unpublished study data were made available to Makris et al.[25] These unpublished data were not made publically available and thus not available for evaluation here. However, even if such information were made publicly available, use of such clarifying information from this study without attempts to contact other study authors to clarify uncertainties in other studies is a direct form of bias in the conduct of an SR and thus is viewed as unfavorable here.

Additional challenges in the integration of RoB are associated with use of RoB alone as a measure of data quality. Often regarded as an ambiguous term, OHAT addressed the role of RoB as part of an evaluation of data quality, noting that internal validity (RoB), external validity (directness), and completeness in reporting are all important elements of assessing the credibility of individual studies.[2] Historically, in practice, other systems such as Klimisch scoring[82] have been implemented. In such systems, guideline-based studies conducted via good laboratory practice (GLP) are regarded as the top quality or "gold standard" studies. A commonly discussed challenge in the uptake of a question-based RoB approach is that these "gold standard" studies do not automatically rank highest. In the context of SR, the elements of a guideline-based or GLP study are not all addressed by RoB, but rather by integration of other components. Many aspects of these "fit for purpose" studies are evaluated as directness or external validity and/or are addressed at the level of inclusion/exclusion (ie, only direct or "fit for purpose" studies would be included in a review). Here, each study was evaluated both for internal and external validity. The guideline/GLP study (Carney et al[67]) and guideline-type studies (ie, experiments conducted following protocols similar to guideline studies, as opposed to hypothesis generating, research-oriented protocols; Schwetz et al,[53] Hardin et al,[54] Healy et al[69]) received more favorable RoB ratings and also higher ratings for directness—the combination of which increase confidence in the outcomes of these higher quality studies.

An example of the challenge in using RoB to critically appraise guideline-based studies (and a recognized shortcoming of this assessment) is accounting for the number of animals in each study (ie, "n"). One of the many components addressed in any given study guideline is that the "n" per dose group should be large enough to capture a potential effect. The OHAT RoB questions do not directly address this. For example, in the TCE-CHD case study, most of the experimental animal studies involving oral exposure (including Johnson et al[20]) did not include adequate animal numbers based on the OECD guideline protocol for developmental toxicology[34] (most included $n < 20$), whereas the majority of the inhalation studies met or exceeded this guideline standard ($n \geq 20$). Although this aspect would indirectly relate to selection, performance, detection, and other (statistical) bias, it was not directly accounted for in the RoB here. Rather than a reflection of study quality per se, this element relates to study sensitivity; high potency chemical effects may still be detected in studies with less than optimal "n" and are more of a design limitation for studies reporting negative data (ie, Were there enough animals per group to capture low potency chemical effects?). This study design element would have further differentiated the oral and inhalation evidence streams within the experimental animal evidence base. In future refinements of critical appraisal tools, this aspect could be added as a subdomain or as a completely separate RoB question. It is thus notable, and commendable, that initial information available regarding updates to the IRIS program suggest that in the future, individual studies will be evaluated for study sensitivity, that is, the ability of the study to detect the potential effect in question[83]; assessment of such would likely cover the study "n" as well as other study design elements that may be unique to a given end point.

Additionally, although the NTP OHAT RoB tool has a clear application to human and experimental animal studies, it does not provide guidance on the evaluation of mechanistic data. As such, we did not evaluate RoB in the avian or in vitro studies included by Makris et al.[25] Although this could be regarded as a shortcoming in the context of hazard assessment, it does not detract from integration of study quality relative to selection of candidate data sets. Although the avian and in vitro studies in the TCE evidence base could potentially be useful information for characterizing biological mechanisms underlying cardiac defects,[84-86] they are very indirect in the context of developing toxicity values, particularly when considering the nature of these models relative to the exposure of concern (via pregnant mothers). These studies do not accommodate for the complexity in biological responses versus the human and experimental animal studies, which notably utilized lower exposures (avian and in vitro studies utilized TCE concentrations several orders of magnitude higher than the human and animal studies). In addition, such studies utilize exposure routes that are not relevant (eg, avian models directly injected TCE into the chorioallantoic membrane of the egg[87-92]). Thus, the human and experimental animal studies are more generalizable to population exposures and thus preferred over in vitro and avian data for risk assessment.

In conclusion, we have demonstrated the importance of carrying out the NAS recommendations to assess RoB on studies used as primary data sources for hazard identification and dose–response assessment—a critical element in determining how confidently conclusions can be drawn. This exercise also demonstrates a need for further development and refinement of frameworks to evaluate both internal and external validity for nonhuman studies. It is anticipated that results presented here both (1) provide important information to risk managers regarding the confidence (and uncertainty) in the TCE-CHD evidence base and (2) provide a demonstration of the role of RoB in the development of toxicity values.

Author Contributions

D. Wikoff contributed to conception and design, contributed to analysis and interpretation, drafted the manuscript, and critically revised manuscript. J. D. Urban contributed to design, contributed to acquisition, analysis, and interpretation, drafted the manuscript, and critically revised the manuscript. S. Harvey contributed to design, contributed to acquisition, analysis, and interpretation, drafted the manuscript, and critically revised the manuscript. L. C. Haws contributed to conception and design, contributed to analysis and interpretation, and critically revised the manuscript. All authors gave final approval and agree to be accountable for all aspects of work ensuring integrity and accuracy.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The authors are employed by ToxStrategies, Inc, which is a consulting firm providing services to private and public organizations on toxicology and risk assessment issues. ToxStrategies received consulting fees for research and preparation of the manuscript. L. H. has consulted and given presentations to regulatory agencies and at scientific conferences regarding TCE risk assessment issues.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by ENTEK International. The funders were given the opportunity to review the draft findings; the purpose of such review was to allow input on the clarity of the science presented but not on interpretation of the research findings. The researchers' scientific conclusions and professional judgments were not subject to the funders' control; the contents of this manuscript reflect solely the view of the authors.

ORCID iD

Daniele Wikoff http://orcid.org/0000-0003-0785-161X

Supplemental Material

Supplementary material for this article is available online.

References

1. Office of Health Assessment and Translation. Handbook for conducting a literature-based health assessment using OHAT approach for systematic review and evidence integration. Office of Health Assessment and Translation. Research Triangle Park, NC: Division of the National Toxicology Program. National Institute of Environmental Health Sciences; 2015. Available at: https://ntp.niehs.nih.gov/pubhealth/hat/review/index-2.html. Updated October 02, 2017. Accessed December 29, 2017.
2. Rooney AA, Boyles AL, Wolfe MS, Bucher JR, Thayer KA. Systematic review and evidence integration for literature-based environmental health science assessments. *Environ Health Perspect*. 2014;122(7):711-718.
3. Stephens ML, Betts K, Beck NB, et al. The emergence of systematic review in toxicology. *Toxicol Sci*. 2016;152(1):10-16.
4. European Food Safety Authority. *Tools for Critically Appraising Different Study Designs, Systematic Review and Literature Searches*. European Food Safety Authority Supporting Publication; 2015. Technical Report EN-836; published July 1, 2015.
5. National Academies of Sciences. *Review of EPA'S Integrated Risk Information System (IRIS) Process*. Washington, DC: National Research Council, The National Academies Press; 2014.
6. National Academies of Sciences. *Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde*. Washington, DC: National Research Council, The National Academies Press; 2011.
7. Eden J, Levit L, Berg A, Morton SC. *Finding What Works in Health Care: Standards for Systematic Reviews. National Academy of Sciences*; 2011. Washington DC.
8. Guyatt GH, Oxman AD, Kunz R, et al; GRADE Working Group. Grade guidelines: 8. Rating the quality of evidence—indirectness. *J Clin Epidemiol*. 2011;64(12):1303-1310.
9. Whaley P, Halsall C, Agerstrand M, et al. Implementing systematic review techniques in chemical risk assessment: challenges, opportunities and recommendations. *Environ Int*. 2016;92-93:556-564.
10. Thayer KA, Wolfe MS, Rooney AA, Boyles AL, Bucher JR, Birnbaum LS. Intersection of systematic review methodology with the NIH reproducibility initiative. *Environ Health Perspect*. 2014;122(7):A176-A177.
11. Office of Health Assessment and Translation. *OHAT Risk of Bias Tool for Human and Animal Studies. Office of Health Assessment and Translation*. Research Triangle Park, NC: Division of the National Toxicology Program, National Institute of Environmental Health Sciences; 2015. https://ntp.niehs.nih.gov/pubhealth/hat/review/index-2.html. Accessed December 29, 2017.
12. Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW. SYRCLE's risk of bias tool for animal studies. *BMC Med Res Methodol*. 2014;14:43.
13. Morgan RL, Thayer KA, Bero L, et al. Grade: assessing the quality of evidence in environmental and occupational health. *Environ Int*. 2016;92-93:611-616.

14. Woodruff TJ, Sutton P. The navigation guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. *Environ Health Perspect*. 2014;122(10):1007-1014.
15. United States Environmental Protection Agency. Procedures for Prioritization of Chemicals for Risk Evaluation Under the Toxic Substances Control Act. 2017. United States Environmental Protection Agency. Procedures for Prioritization of Chemicals for Risk Evaluation Under the Toxic Substances Control Act. EPA-HQ-OPPT-2016-0636, Sept. 18, 2017. https://www.federalregister.gov/documents/2017/07/20/2017-14325/procedures-for-prioritization-of-chemicals-for-risk-evaluation-under-the-toxic-substances-control. Accessed December 29, 2017.
16. Hardin BD, Kelman BJ, Brent RL. Trichloroethylene and cardiac malformations. *Environ Health Perspect*. 2004;112(11):A607-A608; author reply A608-A609.
17. Hardin BD, Kelman BJ, Brent RL. Trichloroethylene and dichloroethylene: a critical review of teratogenicity. *Birth Defects Res A Clin Mol Teratol*. 2005;73(12):931-955.
18. Watson RE, Jacobson CF, Williams AL, Howard WB, DeSesso JM. Trichloroethylene-contaminated drinking water and congenital heart defects: a critical analysis of the literature. *Reprod Toxicol*. 2006;21(2):117-147.
19. DeSesso JM, Risotto SP. Review of TCE cardiac defects data by Makris et al. is not systematic. *Reprod Toxicol*. 2017;71:134.
20. Johnson PD, Goldberg SJ, Mays MZ, Dawson BV. Threshold of trichloroethylene contamination in maternal drinking waters affecting fetal heart development in the rat. *Environ Health Perspect*. 2003;111(3):289-292.
21. United States Environmental Protection Agency. *Toxicological Review of Trichloroacetic Acid (casrn 76-03-9) in Support of Summary Information on the Integrated Risk Information System (IRIS) [EPA report]*. (EPA/635/r-09/003f), Washington, DC. 2011.
22. National Research Council. *Assessing the Human Health Risks of Trichloroethylene: Key Scientific Issues*. Washington, DC: The National Academies Press; 2006.
23. Scientific Committee on Occupational Exposure Limits. *Recommendation from the Scientific Committee on Occupational Exposure Limits for Trichloroethylene*. European Commission. Scoel/sum/142. 2009. http://ec.europa.eu/social/main.jsp?catId=148&intPageId=684&langId=en. Accessed December 29, 2017. Modified April 11, 2017.
24. California EPA's Office of Environmental Health Hazard Assessment. Public health goal for trichloroethylene in drinking water. Prepared by Pesticide and Environmental Toxicology Branch, Office of Environmental Health Hazard Assessment, California Environmental Protection Agency. 2009. https://oehha.ca.gov/water/public-health-goal/public-health-goal-trichloroethylene-drinking-water. Accessed December 29, 2017.
25. Makris SL, Scott CS, Fox J, et al. A systematic evaluation of the potential effects of trichloroethylene exposure on cardiac development. *Reprod Toxicol*. 2016;65:321-358.
26. Bukowski J. Critical review of the epidemiologic literature regarding the association between congenital heart defects and exposure to trichloroethylene. *Crit Rev Toxicol*. 2014;44(7):581-589.
27. Tyl RW, Marr MC. Developmental toxicity testing—methodology. In: Hood RD, ed. *Developmental and Reproductive Toxicology: A Practical Approach. 3rd ed*. Boca Raton, FL: CRC Press; 2012.
28. Staples RE. Detection of visceral alterations in mammalian foetuses. *Teratology*. 1974;9(3):37-38.
29. Stuckhardt JL, Poppe SM. Fresh visceral examination of rat and rabbit fetuses used in teratogenicity testing. *Teratog Carcinog Mutagen*. 1984;4(2):181-188.
30. Wilson JG. Embryological considerations in teratology. *Ann N Y Acad Sci*. 1965;123:219-227.
31. Christian M. Test methods for assessing female reproductive and developmental toxicology. In: *Principles and Methods of Toxicology*. 5th ed. New York, NY: Publ Informa Healthcare; 2008.
32. Kang YJ ZL, Manson JM. Strain differences in response of sprague-dawley and long evans hooded rats to the teratogen nitrofen. *Teratology*. 1986;34(2):213-223.
33. United States Environmental Protection Agency. Guidelines for developmental toxicity risk assessment. EPA/600/FR-91/001, Risk Assessment Forum, U.S. Environmental Protection Agency, Washington, DC: EPA; 1991.
34. Organisation for Economic Co-operation and Development. OECD guideline for the testing of chemicals: prenatal developmental toxicity study. OECD 414. Adopted January 22, 2001. http://dx.doi.org/10.1787/9789264070820-en. Accessed December 29, 2017.
35. Guzelian PS, Victoroff MS, Halmes NC, James RC, Guzelian CP. Evidence-based toxicology: a comprehensive framework for causation. *Hum Exp Toxicol*. 2005;24(4):161-201.
36. Shepard TH. "Proof" of human teratogenicity. *Teratology*. 1994;50(2):97-98.
37. Shepard TH, Lemire RJ. *Catalog of Teratogenic Agents*. 11th ed. Baltimore, MD and London, England: The Johns Hopkins University Press; 2004.
38. Woodward M. *Epidemiology: Study Design and Data Analysis*. 2nd ed. Boca Raton, FL: CRC Press; 2005.
39. Centers for Disease Control and Prevention. Tobacco use and pregnancy: how does smoking during pregnancy harm my health and my baby? 2017. https://www.cdc.gov/reproductivehealth/maternalinfanthealth/tobaccousepregnancy/index.htm. Accessed September 29, 2017.
40. United States Surgeon General. The health consequences of smoking-50 years of progress: a report of the surgeon general. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health; 2014. https://www.surgeongeneral.gov/library/reports/50-years-of-progress/index.html. Accessed December 29, 2017.
41. Oliveira CI, Fett-Conte AC. Birth defects: risk factors and consequences. *J Pediatr Genet*. 2013;2(2):85-90.
42. Riley EP, Infante MA, Warren KR. Fetal alcohol spectrum disorders: an overview. *Neuropsychol Rev*. 2011;21(2):73-80.

43. Feng Y, Yu D, Yang L, et al. Maternal lifestyle factors in pregnancy and congenital heart defects in offspring: review of the current evidence. *Ital J Pediatr*. 2014;40:85.
44. Grewal J, Carmichael SL, Ma C, Lammer EJ, Shaw GM. Maternal periconceptional smoking and alcohol consumption and risk for select congenital anomalies. *Birth Defects Res A Clin Mol Teratol*. 2008;82(7):519-526.
45. Ramakrishnan A, Lee LJ, Mitchell LE, Agopian AJ. Maternal hypertension during pregnancy and the risk of congenital heart defects in offspring: a systematic review and meta-analysis. *Pediatr Cardiol*. 2015;36(7):1442-1451.
46. Yang J, Qiu H, Qu P, Zhang R, Zeng L, Yan H. Prenatal alcohol exposure and congenital heart defects: a meta-analysis. *PLoS One*. 2015;10(6):e0130681.
47. Zhang D, Cui H, Zhang L, Huang Y, Zhu J, Li X. Is maternal smoking during pregnancy associated with an increased risk of congenital heart defects among offspring? A systematic review and meta-analysis of observational studies. *J Matern Fetal Neonatal Med*. 2017;30(6):645-657.
48. Tola S, Vilhunen R, Jarvinen E, Korkala ML. A cohort study on workers exposed to trichloroethylene. *J Occup Med*. 1980;22(11): 737-740.
49. Brender JD, Shinde MU, Zhan FB, Gong X, Langlois PH. Maternal residential proximity to chlorinated solvent emissions and birth defects in offspring: a case-control study. *Environ Health*. 2014;13:96.
50. Gilboa SM, Desrosiers TA, Lawson C, et al. Association between maternal occupational exposure to organic solvents and congenital heart defects, national birth defects prevention study, 1997-2002. *Occup Environ Med*. 2012;69(9):628-635.
51. United States Environmental Protection Agency. *Toxicological Review of Trichloroethylene (casrn 79-01-6) in Support of Summary Information on the Integrated Risk Information System (iris) [EPA report]*. (EPA/635/r-09/011f), Washington, DC. 2011.
52. United States Environmental Protection Agency. *Tce Developmental Cardiac Toxicity Assessment Update*. Washington, DC: United States Environmental Protection Agency; 2014.
53. Schwetz BA, Leong KJ, Gehring PJ. The effect of maternally inhaled trichloroethylene, perchloroethylene, methyl chloroform, and methylene chloride on embryonal and fetal development in mice and rats. *Toxicol Appl Pharmacol*. 1975;32(1):84-96.
54. Hardin BD, Bond GP, Sikov MR, Andrew FD, Beliles RP, Niemeier RW. Testing of selected workplace chemicals for teratogenic potential. *Scand J Work Environ Health*. 1981;7(suppl 4):66-75.
55. Dawson BV, Johnson PD, Goldberg SJ, Ulreich JB. Cardiac teratogenesis of halogenated hydrocarbon-contaminated drinking water. *J Am Coll Cardiol*. 1993;21(6):1466-1472.
56. Johnson PD, Dawson BV, Goldberg SJ, Mays MZ. Trichloroethylene and cardiac malformations: johnson et al.'s response [letter]. *Environ Health Perspect*. 2004;112(11): A607-A608.
57. Johnson PD, Goldberg SJ, Mays MZ, Dawson BV. Erratum: threshold of trichloroethylene contamination in maternal drinking waters affecting fetal heart development in the rat [erratum]. *Environ Health Perspect*. 2005;113(1):A18.
58. Johnson PD, Goldberg SJ, Mays MZ, Dawson BV. Erratum: erratum for johnson et al. [environ health perspect 113: A18 (2005)]. *Environ Health Perspect*. 2014;1224:(A94).
59. Bove FJ. Public drinking water contamination and birthweight, prematurity, fetal deaths, and birth defects. *Toxicol Ind Health*. 1996;12(2):255-266.
60. Bove FJ, Fulcomer MC, Klotz JB, Esmart J, Dufficy EM, Savrin JE. Public drinking water contamination and birth outcomes. *Am J Epidemiol*. 1995;141(9):850-862.
61. Lagakos SW, Wessen BJ, Zelen M. An analysis of contaminated well water and health effects in Woburn, Massachusetts. *J Am Stat Assoc*. 1986;81(395):583-596.
62. Massachusetts Department of Public Health (MDPH). Woburn environment and birth study (summary). [online]. 1998. http://www.mass.gov/eohhs/docs/dph/environmental/investigations/woburn/woburn-summary-environment-birth-study.Pdf. Accessed August 15, 2017. Modified July 24, 2016.
63. Cosby NC, Dukelow WR. Toxicology of maternally ingested trichloroethylene (tce) on embryonal and fetal development in mice and of tce metabolites on in vitro fertilization. *Fundam Appl Toxicol*. 1992;19(2):268-274.
64. Fisher JW, Channel SR, Eggers JS, et al. Trichloroethylene, trichloroacetic acid, and dichloroacetic acid: do they affect fetal rat heart development? *Int J Toxicol*. 2001;20(5):257-267.
65. Narotsky MG, Weller EA, Chinchilli VM, Kavlock RJ. Nonadditive developmental toxicity in mixtures of trichloroethylene, di(2-ethylhexyl) phthalate, and heptachlor in a 5 x 5 x 5 design. *Fundam Appl Toxicol*. 1995;27(2):203-216.
66. Narotsky MG, Kavlock RJ. A multidisciplinary approach to toxicological screening: ii. Developmental toxicity. *J Toxicol Environ Health*. 1995;45(2):145-171.
67. Carney EW, Thorsrud BA, Dugard PH, Zablotny CL. Developmental toxicity studies in crl: cd (sd) rats following inhalation exposure to trichloroethylene and perchloroethylene. *Birth Defects Res B Dev Reprod Toxicol*. 2006;77(5): 405-412.
68. Dorfmueller MA, Henne SP, York RG, Bornschein RL, Manson JM. Evaluation of teratogenicity and behavioral toxicity with inhalation exposure of maternal rats to trichloroethylene. *Toxicology*. 1979;14(2):153-166.
69. Healy TE, Poole TR, Hopper A. Rat fetal development and maternal exposure to trichloroethylene 100 p.p.m. *Br J Anaesth*. 1982; 54(3):337-341.
70. Yauck JS, Malloy ME, Blair K, Simpson PM, McCarver DG. Proximity of residence to trichloroethylene-emitting sites and increased risk of offspring congenital heart defects among older women. *Birth Defects Res A Clin Mol Teratol*. 2004;70(10): 808-814.
71. Forand SP, Lewis-Michl EL, Gomez MI. Adverse birth outcomes and maternal exposure to trichloroethylene and tetrachloroethylene through soil vapor intrusion in new york state. *Environ Health Perspect*. 2012;120(4):616-621.
72. Goldberg SJ, Lebowitz MD, Graver EJ, Hicks S. An association of human congenital cardiac malformations and drinking water contaminants. *J Am Coll Cardiol*. 1990;16(1):155-164.
73. Ruckart PZ, Bove FJ, Maslia M. Evaluation of exposure to contaminated drinking water and specific birth defects and childhood cancers at marine corps base Camp Lejeune, North Carolina: a case-control study. *Environ Health*. 2013;12:104.

74. DeSesso JM, Venkat AG. Cardiovascular development and malformation. In: Kapp RW, Tyl RW, eds. *Reproductive Toxicology*. 3rd ed. Boca Raton, FL: CRC Press, Taylor & Francis Group; 2010:223-248.
75. Liu Z, Yu Y, Li X, et al. Maternal lead exposure and risk of congenital heart defects occurrence in offspring. *Reprod Toxicol*. 2015;51:1-6.
76. Cox LA Jr, Popken DA, Berman DW. Causal versus spurious spatial exposure-response associations in health risk analysis. *Crit Rev Toxicol*. 2013;43(suppl 1):26-38.
77. Johnson PD, Dawson BV, Goldberg SJ. A review: trichloroethylene metabolites: potential cardiac teratogens. *Environ Health Perspect*. 1998;106(suppl 4):995-999.
78. Epstein DL, Nolen GA, Randall JL, et al. Cardiopathic effects of dichloroacetate in the fetal long-evans rat. *Teratology*. 1992; 46(3):225-235.
79. Smith MK, Randall JL, Read EJ, Stober JA. Teratogenic activity of trichloroacetic acid in the rat. *Teratology*. 1989;40(5): 445-451.
80. Smith MK, Randall JL, Read EJ, Stober JA. Developmental toxicity of dichloroacetate in the rat. *Teratology*. 1992;46(3): 217-223.
81. Johnson PD, Dawson BV, Goldberg SJ. Cardiac teratogenicity of trichloroethylene metabolites. *J Am Coll Cardiol*. 1998;32(2): 540-545.
82. Klimisch HJ, Andreae M, Tillmann U. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul Toxicol Pharmacol*. 1997;25(1):1-5.
83. Bahadori T. *Iris Today—An Update on Progress*. Paper presented at: EPA Science Advisory Board (SAB) Chemical Assessment Advisory Committee (CAAC); September 27-28, 2017.
84. Hunter ES 3rd, Rogers EH, Schmid JE, Richard A. Comparative effects of haloacetic acids in whole embryo culture. *Teratology*. 1996;54(2):57-64.
85. Mishima N, Hoffman S, Hill EG, Krug EL. Chick embryos exposed to trichloroethylene in an ex ovo culture model show selective defects in early endocardial cushion tissue formation. *Birth Defects Res A Clin Mol Teratol*. 2006;76(7):517-527.
86. Saillenfait AM, Langonne I, Sabate JP. Developmental toxicity of trichloroethylene, tetrachloroethylene and four of their metabolites in rat whole embryo culture. *Arch Toxicol*. 1995;70(2):71-82.
87. Bross G, DiFranceisco D, Desmond ME. The effects of low dosages of trichloroethylene on chick development. *Toxicology*. 1983;28(4):283-294.
88. Drake VJ, Koprowski SL, Hu N, Smith SM, Lough J. Cardiogenic effects of trichloroethylene and trichloroacetic acid following exposure during heart specification of avian development. *Toxicol Sci*. 2006;94(1):153-162.
89. Drake VJ, Koprowski SL, Lough J, Hu N, Smith SM. Trichloroethylene exposure during cardiac valvuloseptal morphogenesis alters cushion formation and cardiac hemodynamics in the avian embryo. *Environ Health Perspect*. 2006;114(6):842-847.
90. Elovaara E, Hemminki K, Vainio H. Effects of methylene chloride, trichloroethane, trichloroethylene, tetrachloroethylene and toluene on the development of chick embryos. *Toxicology*. 1979;12(2):111-119.
91. Loeber CP, Hendrix MJ, Diez De Pinos S, Goldberg SJ. Trichloroethylene: a cardiac teratogen in developing chick embryos. *Pediatr Res*. 1988;24(6):740-744.
92. Rufer ES, Hacker TA, Flentke GR, et al. Altered cardiac function and ventricular septal defect in avian embryos exposed to low-dose trichloroethylene. *Toxicol Sci*. 2010;113(2):444-452.